



**Alibaba Group**  
阿里巴巴集团

# SDM: SEQUENTIAL DEEP MATCHING MODEL FOR ONLINE LARGE-SCALE RECOMMENDER SYSTEM

Fuyu Lv Taiwei Jin Changlong Yu Fei Sun  
Quan Lin Keping Yang Wilfred Ng



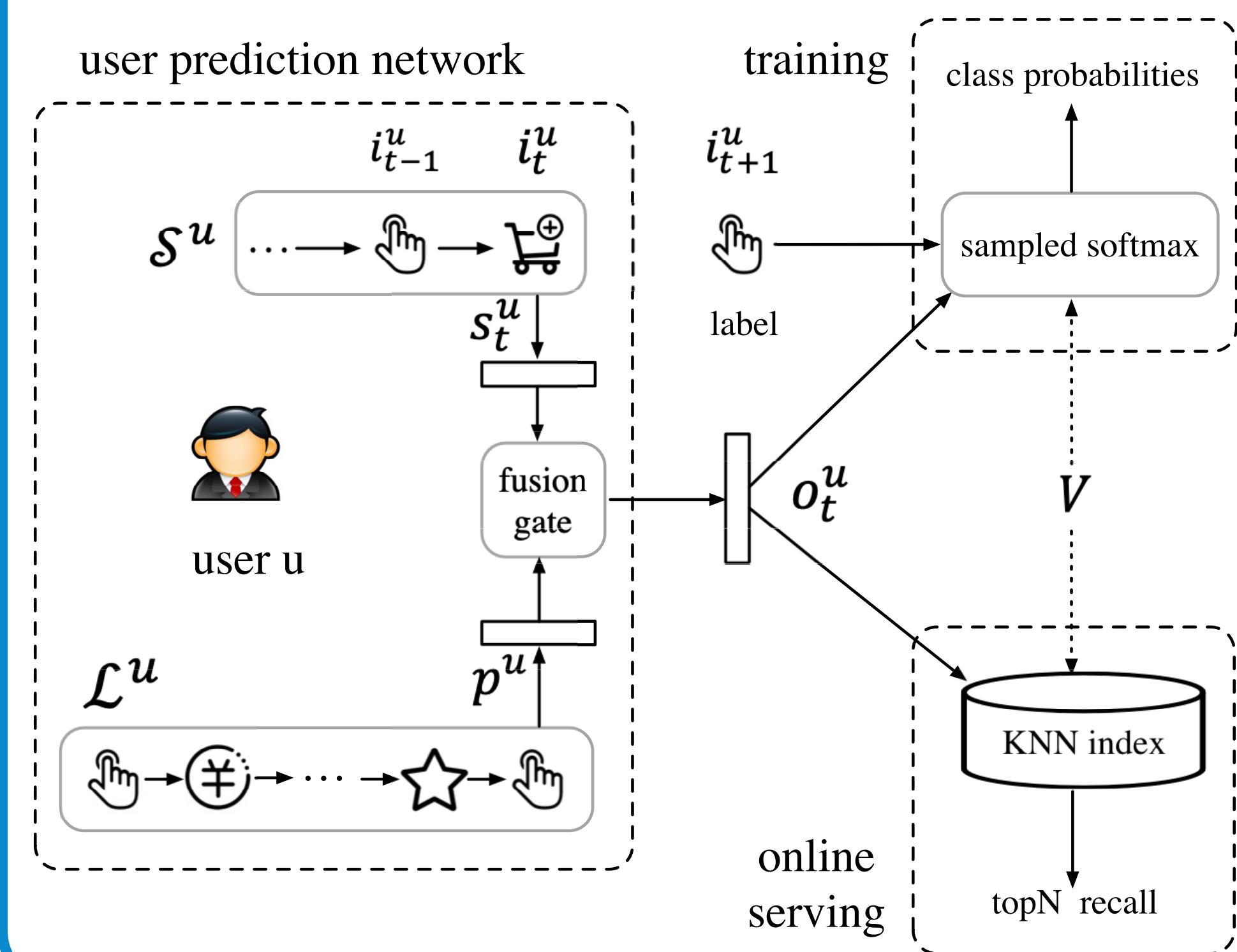
THE HONG KONG  
UNIVERSITY OF SCIENCE  
AND TECHNOLOGY

## MOTIVATION

- To capture dynamic evolution of users' interests, we introduce deep sequential recommendation model in matching stage of recommender system at Taobao.
- For the purpose of representing difference levels of interests, we model user behavior sequences by combining short-term sessions and long-term behaviors.
- We tackle two inherent problems in real-world applications: (1) users' multiple interest tendencies in one session. (2) effectively combining short-term and long-term preferences.

## SYSTEM DEPLOYMENT

The item embedding vectors  $V$  are imported into a KNN searching system. Meanwhile, the user prediction network is deployed on a high-performance real-time system for fast inference.

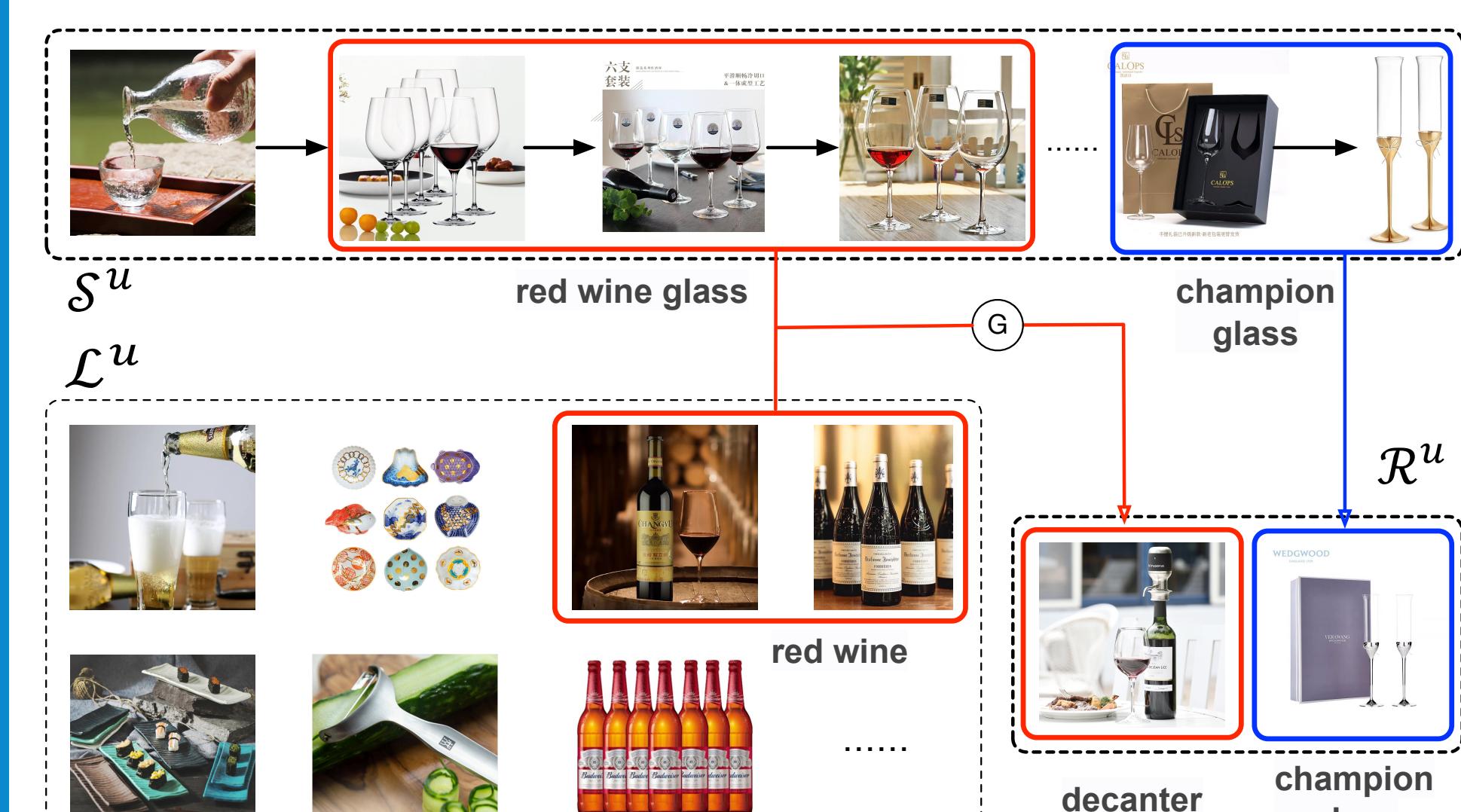


## CASE STUDY

Visualization of attention weights from head<sub>1</sub> to head<sub>4</sub> in multi-head attention over a short-term session  $S^u$  sampled from offline test dataset of Taobao.



A short-term session  $S^u$  and long-term behavior  $L^u$  from a sampled user of our online system.



## SAMPLE SOURCE CODE

We are recommendation team of Alibaba Group.

Welcome to join us!

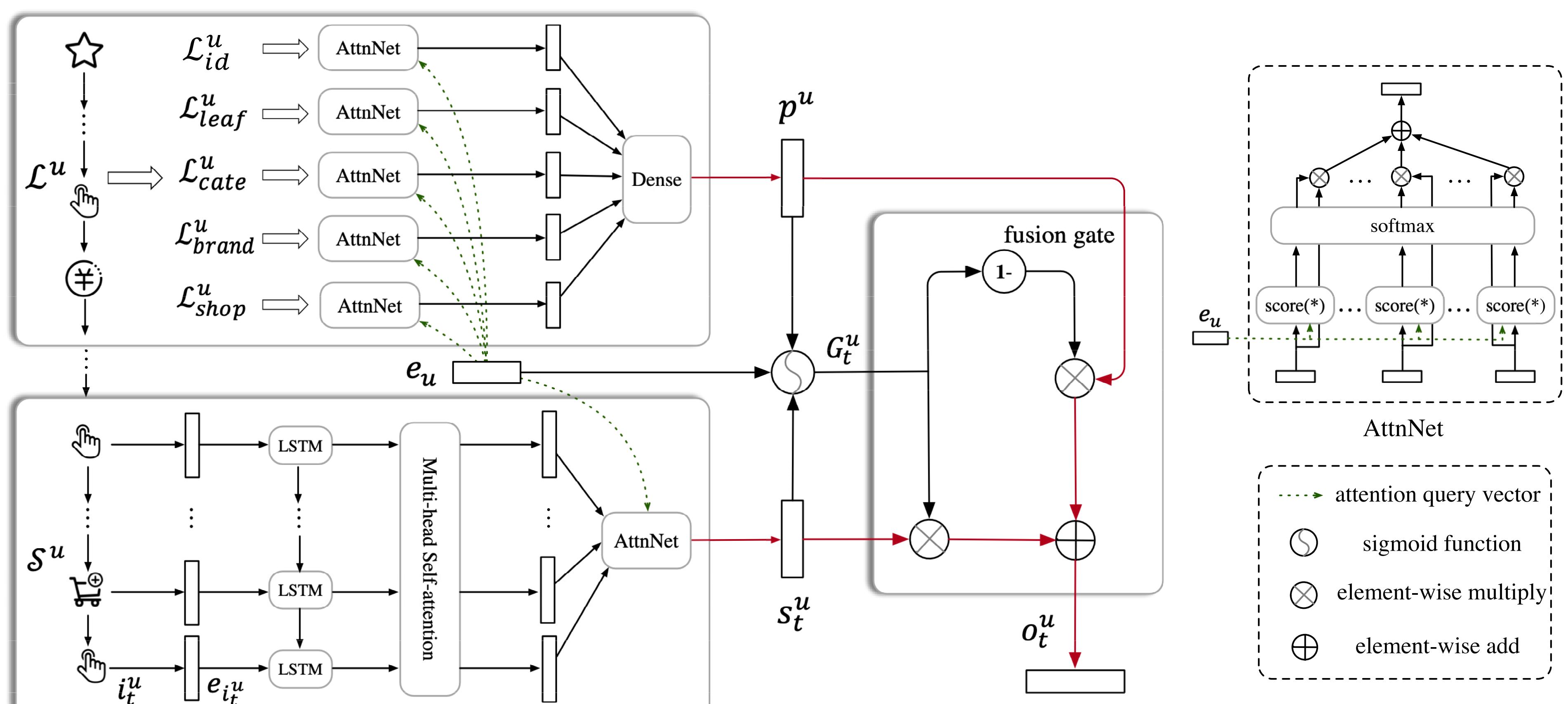


(a) Github



(b) Wechat

## MODEL DESIGN



- Sequence Generation:** We get each latest session of user  $u$ , namely  $S^u$ , by sorting the interacted items. The long-term behaviors of  $u$  that happened before  $S^u$  in past 7 days are denoted by  $L^u$ .
- Embedding:** We describe items and users from different side information e.g, category, brand.
- Recurrent Layer:** To capture and characterize the global temporal dependency in the short-term sequence data, we apply LSTM network as the recurrent cell.
- Attention Mechanism:** Multi-head attention is applied to solve the issue of multiple interests by representing preferences from different views.
- Long-term Behaviors Fusion:** We elaborately design a gated neural network that adopts user profile, short-term and long-term representations as inputs.

## DATA SETS AND EVALUATIONS

### Statistics of Offline and Online Datasets

Dataset	Data Type	Data Split	# <sup>a</sup> Users	#Items	#Records	#Sessions	S.len <sup>b</sup>	L.size <sup>c</sup>	Time Interval
JD	offline	train	802,479	154,568	9,653,777	2,666,189	3.3	20	15/Mar/2018 - 8/Apr/2018
		test	10,366	74,564	498,492	15,069	8.6	20	9/Apr/2018 - 15/Apr/2018
Taobao	offline	train	498,633	2,053,798	45,157,298	7,011,385	6.1	20	15/Dec/2018 - 21/Dec/2018
		test	13,237	588,306	1,170,401	13,237	9.2	20	22/Dec/2018
	online	train	$3.3 \times 10^8$	$1 \times 10^8$	$2.1 \times 10^{10}$	$2.7 \times 10^9$	7.1	50	Dec/2018
	test	$3.3 \times 10^8$	$1 \times 10^8$	/	/	8.1	50	Dec/2018	

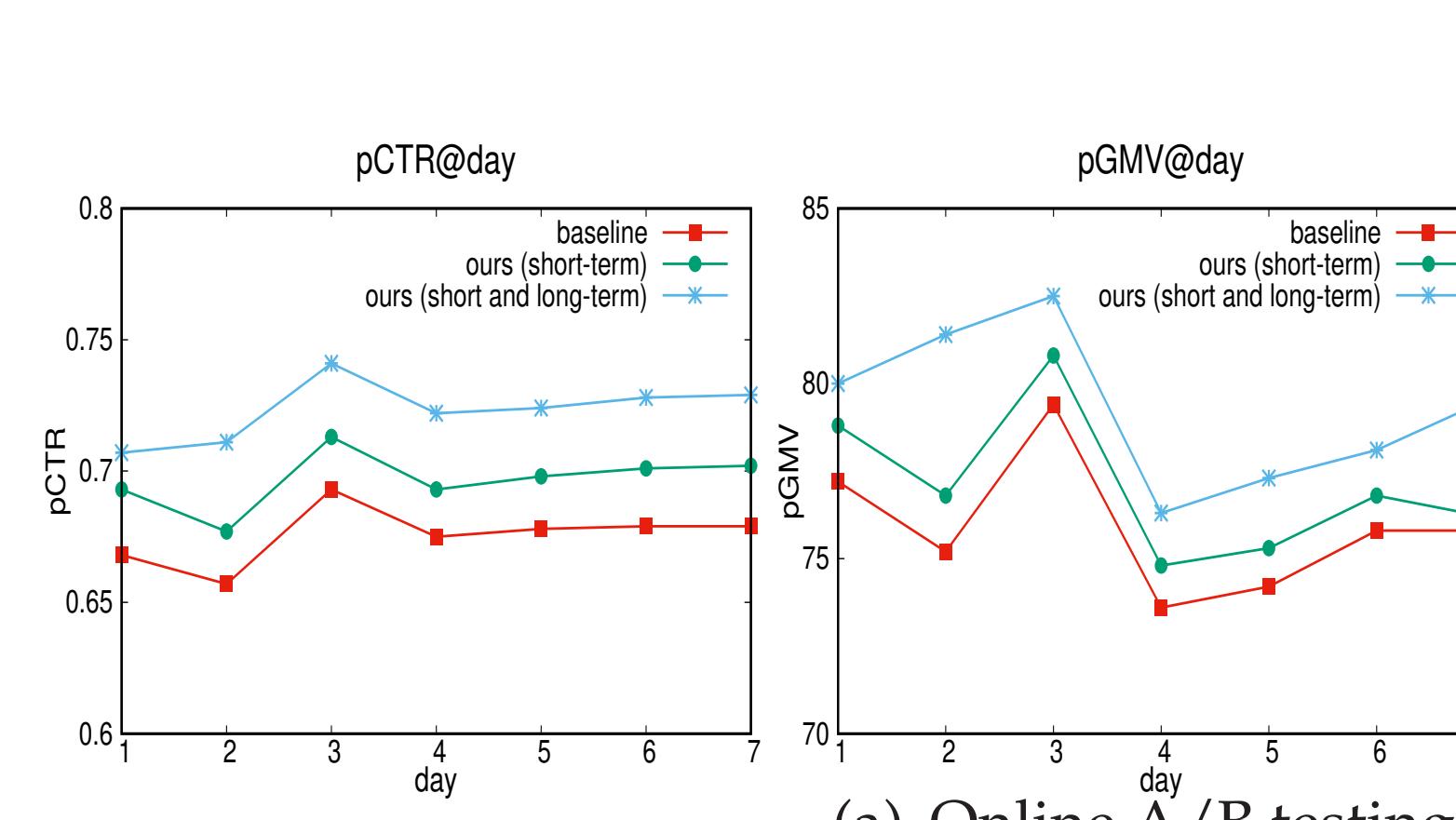
<sup>a</sup># means the number of. <sup>b</sup>S.len is the average length of short-term behaviors. <sup>c</sup>L.size is the maximum size of each subset in long-term behaviors.

### Evaluation Metrics

- For offline testing, we use HitRate@K, Precision@K, Recall@K and F1@K.
- For online testing, we consider the most important online metrics: pCTR, pGMV and discovery.

## EXPERIMENTS

Models	Taobao				JD			
	HitRate@100	Recall@100	Precision@100	F1@100	HitRate@20	Recall@20	Precision@20	F1@20
Item-based CF	60.27%	3.24%	2.00%	2.43%	67.50%	9.08%	9.41%	8.99%
DNN	60.88%	2.85%	1.83%	2.18%	68.43%	8.93%	9.65%	8.98%
GRU4REC	65.60%	3.66%	2.30%	2.77%	69.44%	9.33%	9.83%	9.29%
NARM	66.97%	3.57%	2.25%	2.70%	70.33%	9.07%	9.58%	9.04%
SHAN	67.30%	3.71%	2.33%	2.80%	70.54%	9.42%	10.02%	9.41%
BINN	67.55%	3.49%	2.20%	2.64%	72.19%	9.38%	9.93%	9.36%
SDMMA	68.24%	3.68%	2.32%	2.79%	70.41%	9.21%	9.72%	9.18%
PSDMMA	69.43%	3.75%	2.37%	2.84%	71.21%	9.21%	9.78%	9.20%
PSDMMAL	70.72%	<b>3.86%</b>	2.44%	<b>2.93%</b>	73.25%	9.47%	10.13%	9.48%
PSDMMAL-N	<b>73.13%</b>	3.83%	<b>2.45%</b>	2.92%	<b>74.33%</b>	<b>9.68%</b>	<b>10.42%</b>	<b>9.72%</b>
PSDMMAL-NoS	65.41%	3.38%	2.14%	2.56%	70.07%	9.05%	9.60%	9.03%



(a) Online A/B testing

#heads	HitRate@K	Recall@K	Precision@K	F1@K
1	70.00%	3.82%	2.40%	2.88%
2	70.64%	3.83%	2.41%	2.89%
4	<b>70.72%</b>	<b>3.86%</b>	<b>2.44%</b>	<b>2.93%</b>
8	70.21%	3.77%	2.37%	2.85%

fusion	HitRate@K	Recall@K	Precision@K	F1@K
multiply	67.09%	3.42%	2.16%	2.59%
concat	69.74%	3.70%	2.34%	2.80%
add	70.24%	3.75%	2.37%	2.84%
gated	<b>70.72%</b>	<b>3.86%</b>	<b>2.44%</b>	<b>2.93%</b>

(b) Effectiveness of multi-head and gate